

Appendix

We present additional details and experiments of the proposed CityWalkers dataset and the PedGen model for pedestrian movement generation. In Sec. A, we introduce more details and statistics of CityWalkers. Sec. B benchmarks the noise level of our automatic labeling pipeline. In Sec. C, we discuss implementation details of PedGen. Sec. D provides more visualizations of the dataset and the qualitative results. We discuss important licensing and privacy considerations in Sec. E and broader impacts of our work in Sec. F. More qualitative results can be found in the supplementary video.

A MORE DETAILS ON CITYWALKERS

We manually review all videos to ensure the raw data quality and scene diversity. We collected 728 such videos and split each one into 5-second clips, with an interval of 25 seconds between subsequent clips. We then perform a second check round to examine whether all clips are in urban environments, have proper lighting conditions without much motion blur, and do not contain abrupt viewpoint changes. To prevent the leaking of personally identifiable information, we also blur the faces and license plates with mosaicing tools (Xu et al., 2020). In total, 22,698 clips are collected, each representing a different urban scene. All video clips are decomposed into image sequences with a frame rate of 30 fps. We keep pedestrians tracked consecutively for at least 10 frames, and pass it together with the camera angular velocity predicted by DPVO (Teed et al., 2023) to the WHAM network to get pseudo-labels for 4D global pedestrian movement. For each pedestrian, WHAM outputs both its movement in the global frame $\mathcal{X}^g = \{\mathbf{t}_t^g, \phi_t^g, \theta_t, \beta_t\}_{t=1}^T$, and in the local camera frame $\mathcal{X}^c = \{\mathbf{t}_t^{c(t)}, \phi_t^{c(t)}, \theta_t, \beta_t\}_{t=1}^T$, note that the local camera frame $c(t)$ is varying with time as the camera is also moving in the video. To align the context information in the local camera frame at a specific timestep τ and the future pedestrian movement label in the global frame, we define an additional transformation matrix $\mathbf{T}_g^{c(\tau)} = [\mathbf{R}_g^{c(\tau)} | \mathbf{t}_g^{c(\tau)}]$ with $\mathbf{R}_g^{c(\tau)} = \phi_\tau^{c(\tau)}(\phi_\tau^g)^{-1}$, $\mathbf{t}_g^{c(\tau)} = \mathbf{t}_\tau^{c(\tau)} - \phi_\tau^{c(\tau)}(\phi_\tau^g)^{-1}\mathbf{t}_\tau^g$ and apply $\mathbf{T}_g^{c(\tau)}$ on the global human motion \mathcal{X}^g . The result is the global pedestrian movement in the fixed local camera coordinate $c(\tau)$, which can be paired with the context information as $\{\mathcal{I}_\tau, \mathcal{I}_\tau^d, \mathcal{I}_\tau^s, \mathcal{X}^{c(\tau)}\}$ to get training and validation samples. We use the CityScapes (Cordts et al., 2016) classes for the semantic map, as they contain common classes in urban scenes, such as buildings, sidewalks, and cars.

Table 4 compares CityWalkers to other human motion datasets. CityWalkers has the most diverse human subjects and scenes compared to other human motion datasets and is the only dataset that uses web source videos and pseudo-labels. We provide further statistics regarding the pedestrian movements in CityWalkers in Fig. 6. Plots A-D display key motion characteristics. Plot A shows that our dataset captures motions with a wide range of typical human walking speeds. As evidenced by Plot B and C, our data also contains substantial samples of varying stride patterns. We also demonstrate the diversity of movement directions with Plot D, which represents the change in orientation across the recorded motion. In addition, we plot pedestrian body shape statistics with Plot E and Plot F. We look at the height of pedestrians in Plot E and their waist-to-height ratio in Plot F, as an indicator for the mass index. Fig. 7 demonstrates the list of cities and countries in CityWalkers and its pedestrian attributes roughly estimated by an off-the-shelf VLM (Chen et al., 2023b). CityWalkers covers most European countries and some Asian countries, and we plan to add more locations in the future. As most of the places in CityWalkers have many tourists, its pedestrians are from all over the world, and the age groups and genders are well-represented.

B BENCHMARKING THE NOISE LEVEL OF CITYWALKERS

Though we have applied several techniques to improve the quality of pseudo-labels by using state-of-the-art models and filtering wrong predictions, label noise from web videos is still inevitable and hence it is important to benchmark the accuracy of our data autolabeling pipeline. We use the SLOPER4D dataset (Dai et al., 2023), collected in a similar outdoor setting as CityWalkers with a much smaller scale. The SLOPER4D dataset has ground-truth human motion and scene-depth labels annotated from the 3D LiDAR point clouds. We evaluate the accuracy of the 4D human motion estimation with the Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) and the World-Aligned Mean Per Joint Position Error (WA-MPJPE), and the accuracy of the monocular depth

Table 4: **Comparison of human motion datasets.** We compare CityWalkers with datasets that do not have scene context (top), datasets that provide scene context labels but are captured in controlled environments (middle), and datasets that are captured in the wild (bottom). CityWalkers has the most number of scenes and subjects among all datasets and is the only dataset that uses web source videos and pseudo-labels.

Datasets	Web source	Pseudo-labels	# Subjects	# Scenes	Hours	Depth	Segmentation	SMPL
AMASS (Mahmood et al., 2019)	-	-	344	-	62.9	-	-	✓
DNA-Rendering (Cheng et al., 2023b)	-	-	1,500	-	3.2	-	-	✓
PROX (Hassan et al., 2019)	-	-	20	12	0.9	✓	✓	✓
RELI11D (Yan et al., 2024)	-	-	10	7	3.3	✓	-	✓
RICH (Huang et al., 2022)	-	-	22	5	2.7	✓	-	✓
TRUMANS (Jiang et al., 2024)	-	-	7	100	15.0	✓	✓	✓
3DPW (Von Marcard et al., 2018)	-	-	7	60	0.5	-	-	✓
EMDB (Kaufmann et al., 2023)	-	-	10	81	1.0	-	-	✓
SLOPER4D (Dai et al., 2023)	-	-	12	10	1.4	✓	-	✓
JRDB-Pose (Vendrow et al., 2023)	-	-	5,022	54	1.1	✓	✓	-
CityWalkers (ours)	✓	✓	120,914	16,215	30.8	✓	✓	✓

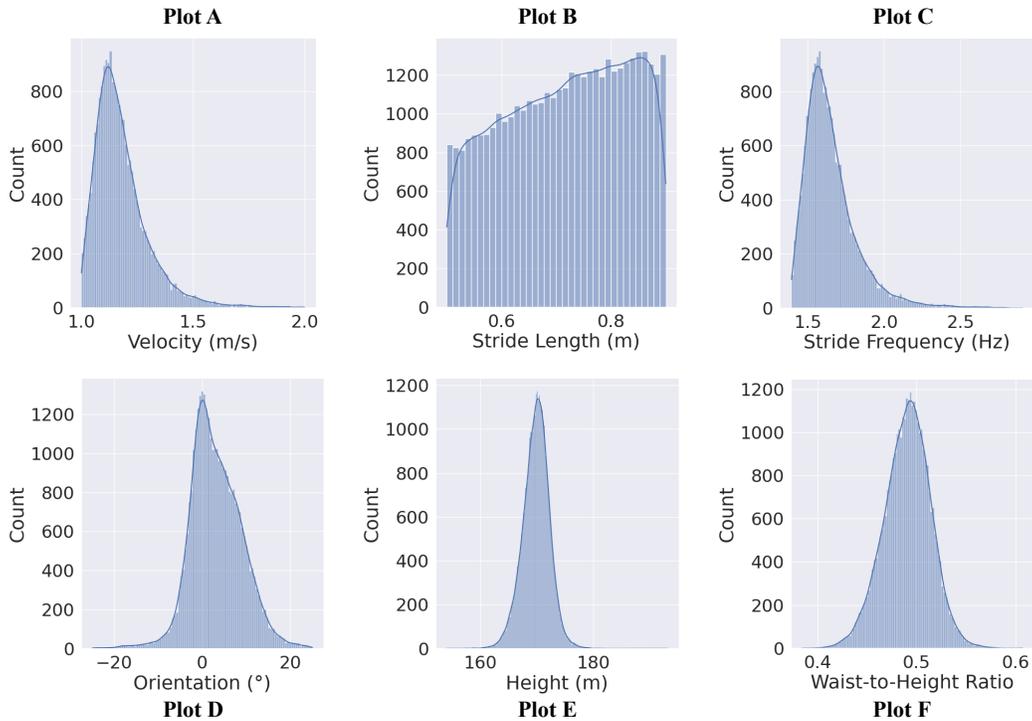


Figure 6: **Pedestrian movement statistics of CityWalkers.**

estimation with the absolute relative error (REL) and the Root Mean Squared Error (RMSE). The results are shown in Tab. 5. We want to stress that while encountering label noise is inevitable, web videos are necessary to learn a generalizable pedestrian movement generation model with natural pedestrian behaviors and diverse motion contexts.

Table 5: **Benchmarking results of our autolabeling pipeline on the SLOPER4D dataset.**

4D Human Motion		Depth	
PA-MPJPE (mm) ↓	WA-MPJPE (mm) ↓	REL ↓	RMSE ↓
42.76	297.72	0.33	5.46

972 according to the depth label. We can see that the ground truth movements align well with the geometry
973 of the scene, showing the quality of the movement and scene context labels. The generated pedestrian
974 movements also match well with their surrounding environments in 3D. Fig. 11 shows a qualitative
975 comparison between PedGen and other baselines. It can be noticed that PedGen can generate more
976 natural and diverse motions with different poses, gestures, and hand movements. It also aligns with
977 the surrounding environment better after being conditioned on the context factors. Some additional
978 4D pedestrian movement labels in CityWalkers are visualized in Fig. 10 to show the diversity of the
979 pedestrian movements. Fig. 12 shows samples of the CARLA test set that we used to evaluate the
980 zero-shot generalization ability of PedGen. We display a handful of urban scenes with diverse objects
981 and layouts, all of which reflect test set diversity. In Fig. 13, we lay out a wide array of scenes in
982 CityWalkers with diverse locations, weather, crowd density, and time of day to show the diversity of
983 the urban scenes. Fig. 14 shows visualizations of the automatically filtered anomaly samples. We can
984 observe that the anomaly labels in the first iteration of anomaly filtering have drastic errors in the
985 body pose or do not belong to pedestrian movements. In contrast, the anomaly labels in the second
986 iteration of anomaly filtering have much smaller errors with minor deviations in the local movements
987 and could also contain false positives with novel movements. We find that two iterations with a
988 reconstruction error threshold of 10 are sufficient to filter most low-quality labels with the best model
989 performance. An ablation study on the number of filtering iterations can be found in Tab. 6.

990 E ETHICS AND PRIVACY CONSIDERATIONS

991
992 We will follow practices in the existing YouTube datasets (Abu-El-Haija et al., 2016; Xu et al., 2018)
993 for privacy protection on web videos. All of our raw videos (Sczepansky, 2024) have a Creative
994 Commons license (CC-BY-SA¹) and are complied with YouTube’s privacy policy and terms of
995 service². Besides, we skip copyright-related information during data processing to protect the rights
996 of logos, channel owner information, or other copyrighted materials. We will not provide processed
997 video clips, and users are redirected to original YouTube videos via a link and follow our provided
998 pre-processing scripts to process the dataset themselves.

999 Our data will be released under the CC BY-NC-SA 4.0 license³. To protect privacy, we remove
1000 all personally identifiable information by adding mosaics to faces and license plates in the videos
1001 following (Grauman et al., 2022). We will also implement safeguards on our dataset webpage with
1002 detailed user agreements and rules to encrypt personal information, enforce access limits, and monitor
1003 misuse and unauthorized data access. Upon data release, we will credit the source, provide a link to
1004 the license, and state that no modifications have been made to the raw videos except for the appended
1005 pedestrian movement and scene labels, and that the data shall not be used for commercial purposes.
1006 We will follow the guidelines of our institute’s institutional review board and comply with applicable
1007 laws. For instance, the human subjects in the videos have the right to view, correct, and delete
1008 personal information in the dataset. The review board will also evaluate the collection, use, and
1009 sharing of the data to ensure alignment with best practices in data privacy and ethics.

1010 F BROADER IMPACTS

1011
1012 Our work can benefit many applications. For example, city designers can simulate pedestrian
1013 movements to optimize public areas and transportation systems. Forecasting future pedestrian
1014 movements is also crucial for the safe deployment of autonomous vehicles. The diverse urban scenes
1015 and pedestrians in the CityWalkers dataset could also support future research directions besides
1016 pedestrian movement generation, such as relation modeling between real-world scenes and humans
1017 and integration of realistic human movements and urban scenes into embodied AI training.

1018
1019 Some potential negative societal impacts of our work include using the released data for surveillance
1020 applications by modeling and predicting pedestrian behaviors and the risk of leaking personally
1021 identifiable information. It is also possible that our model could be misused to generate fake pedestrian
1022 movements of real-world humans. As discussed in Sec. E, we will enforce practices to protect privacy

1023
1024 ¹<https://creativecommons.org/licenses/by/3.0/legalcode.en>

1025 ²<https://www.youtube.com/static?template=terms>

³<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en>

(e.g., removing personally identifiable information) and add the user agreement and the license to our dataset to prevent misuse.

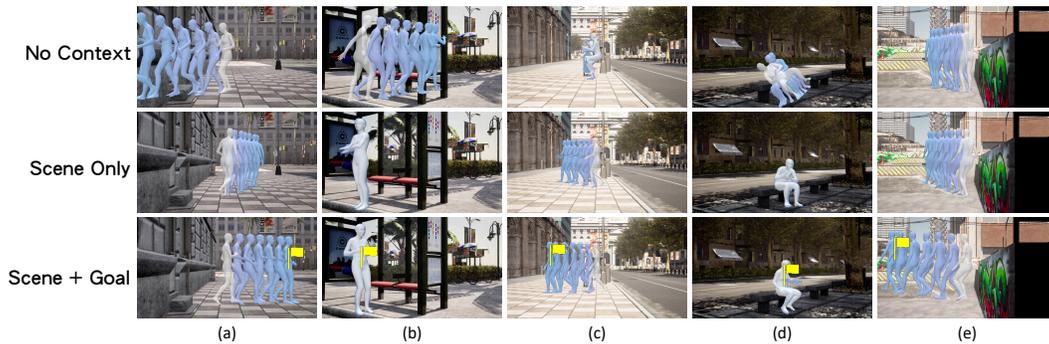


Figure 8: **Qualitative comparisons of the scene and the goal context.** We visualize the generated movements using no context factor, the ones using only the scene context, and the ones using both the scene and the goal context (marked as the yellow flag) in five environments in the CARLA simulator. From the comparison between "No Context" and "Scene Only," we can see that (a) The generated movement with no context hits the wall, while the one with the scene context can navigate the sidewalk. (b): The generated movement with no context directly walks into the bus stop, while the model generates a standing pose with a reasonable direction after using the scene context. (c): The model generates a sitting pose and makes the movement float from the ground if no context is used, and it generates a walking motion while avoiding the obstacles in the front after using the scene context. (d): The sitting pose generated without context factors is jittering and has the wrong orientation, while the generated sitting pose is stable and more plausible after using the scene context. (e): The generated movement ignores the slope in the front without the context. Adding the scene context makes the model aware of the terrain and the movement to walk upward. From the comparison between "Scene Only" and "Scene + Goal", we can see that adding the goal context can make the movement reach the goal more precisely, while in some cases like (b) and (d), using the scene context alone can generate plausible human poses similar to the ones with the goal context.



Figure 9: **Visualizations in 3D.** We visualize both the ground truth scene and movement labels (orange) and the generated movements by PedGen (blue) in multiple views in 3D by unprojecting the image pixels from the depth labels.

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133



Figure 10: **Samples of 4D pedestrian movement labels in CityWalkers.** The text descriptions of the movements from top left to bottom right are: walking down stairs (pink), turning and lifting baggage up steps (light green), walking up stairs (dark purple), turning around with phone in hand (sky blue), moving hands to hip (dark green), wiping seats and tables (red), jumping and skipping around (yellow), taking photo and standing up (light purple).

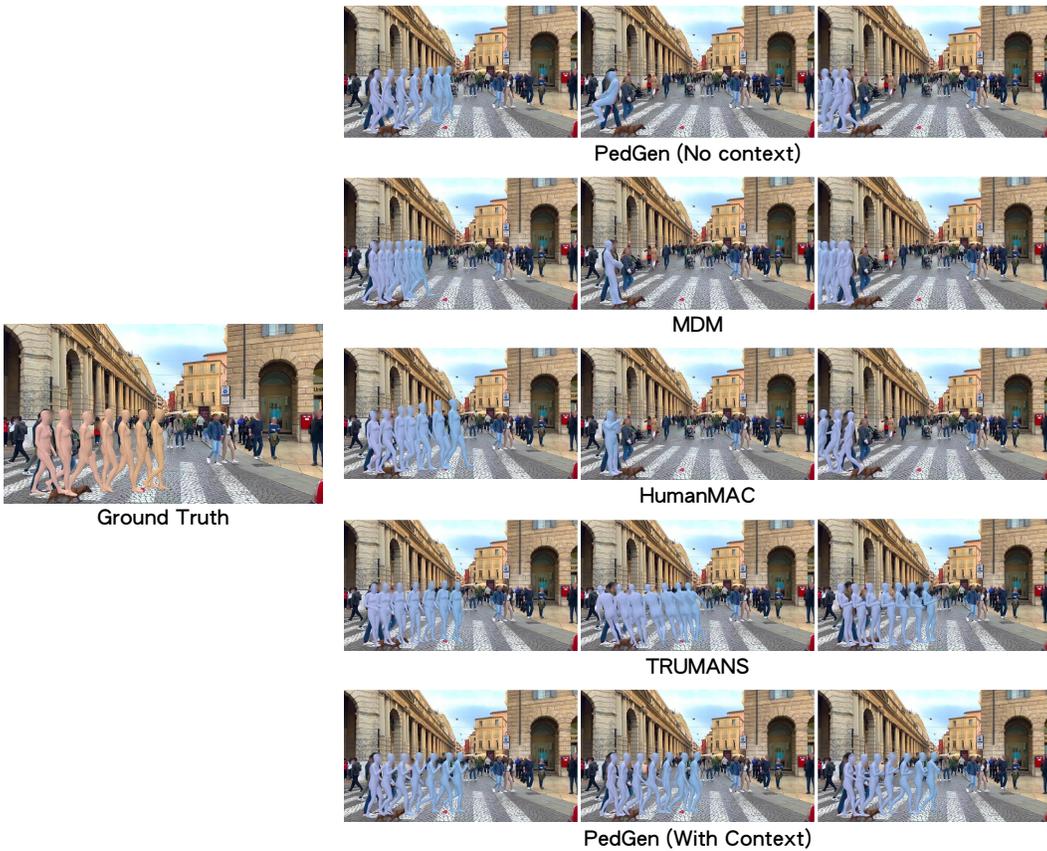


Figure 11: **Qualitative comparison results.** We visualize the generation results of PedGen compared to the other baselines and the ground truth. Three random samples are generated for each method.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

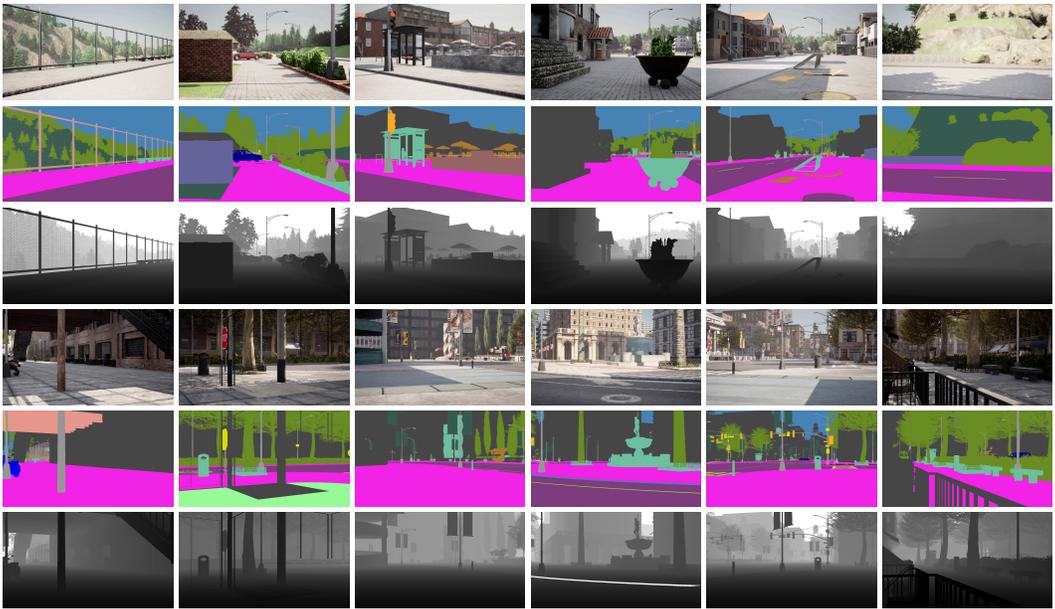


Figure 12: Samples in the CARLA test set. Each scene contains a rendered image, a semantic map, and a depth map.

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

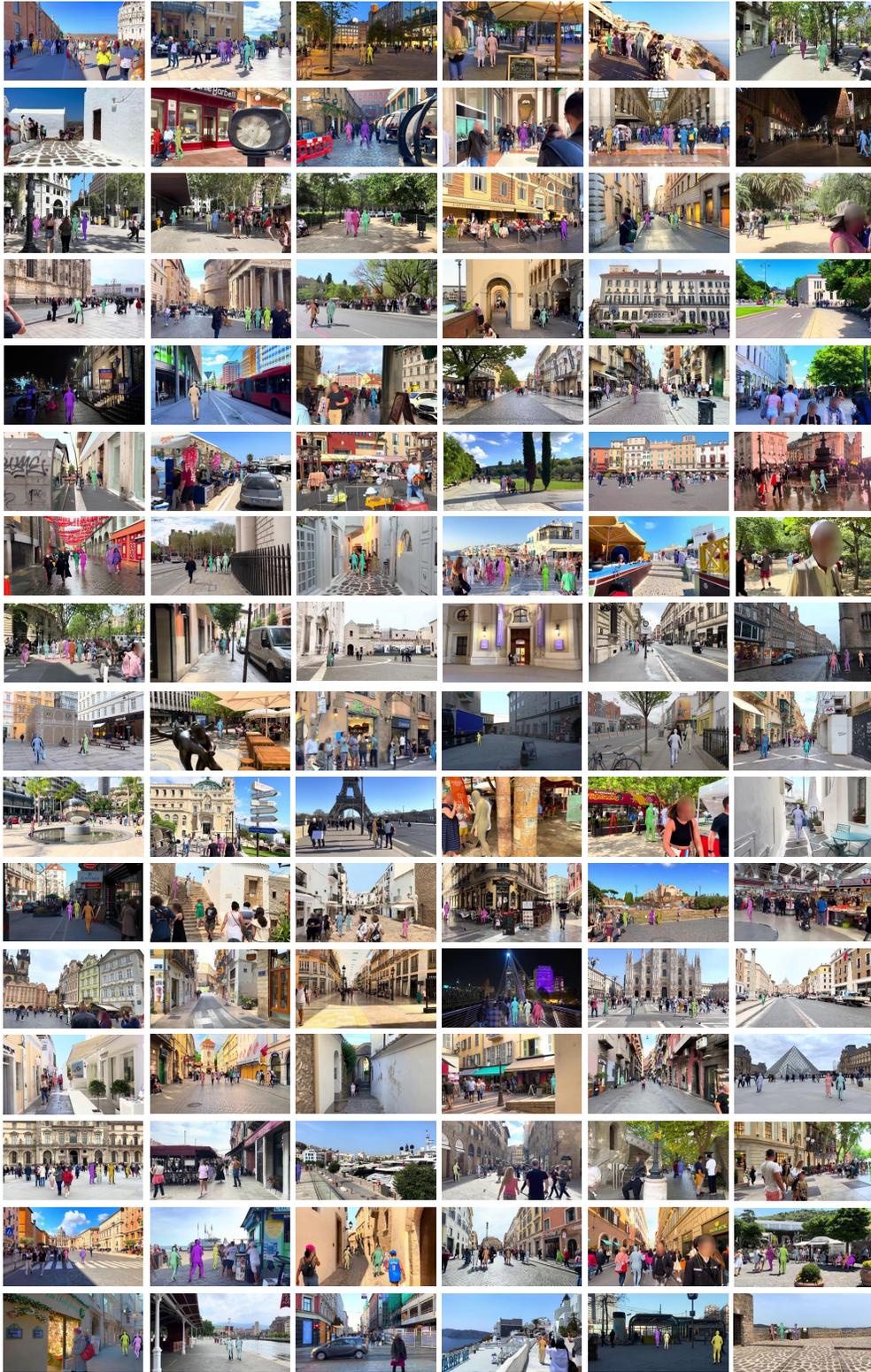


Figure 13: Samples of real-world scenes in CityWalkers. We visualize the extracted pedestrian 3D meshes in each scene.



1253 **Figure 14: Visualizations of anomaly labels in CityWalkers.** We visualize the filtered labels in the first and the
1254 second iterations of automatic anomaly label filtering.

1255
1256 **Table 6: Ablation on the number of filtering iterations.** We evaluate PedGen with no context on the
1257 CityWalkers validation set.

1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

Filtering Iterations	Metric			
	mADE ↓	aADE ↓	mFDE ↓	aFDE ↓
0	1.17	4.45	1.64	8.31
1	1.17	4.22	1.68	7.88
2	1.13	4.32	1.60	8.09
3	1.17	4.42	1.63	8.22